Inherent size constraints on prokaryote gene networks due to "accelerating" growth

M. J. Gagen and J. S. Mattick
ARC Special Research Centre for Functional and Applied Genomics,
*Institute for Molecular Bioscience, University of Queensland, Brisbane, Qld 4072, Australia**
(Dated: February 9, 2008)

Networks exhibiting "accelerating" growth have total link numbers growing faster than linearly with network size and can exhibit transitions from stationary to nonstationary statistics and from random to scale-free to regular statistics at particular critical network sizes. However, if for any reason the network cannot tolerate such gross structural changes then accelerating networks are constrained to have sizes below some critical value. This is of interest as the regulatory gene networks of single celled prokaryotes are characterized by an accelerating quadratic growth and are size constrained to be less than about 10,000 genes encoded in DNA sequence of less than about 10 megabases. This paper presents a probabilistic accelerating network model for prokaryotic gene regulation which closely matches observed statistics by employing two classes of network nodes (regulatory and non-regulatory) and directed links whose inbound heads are exponentially distributed over all nodes and whose outbound tails are preferentially attached to regulatory nodes and described by a scale free distribution. This model explains the observed quadratic growth in regulator number with gene number and predicts an upper prokaryote size limit closely approximating the observed value.

I. INTRODUCTION

The rapidly expanding field of network analysis, reviewed in [1, 2], has provided examples of networks exhibiting "accelerating" network growth, where link number grows faster than linearly with network size. For instance, the Internet [3] appears to grow by adding links more quickly than sites though the relative change over time is small and the Internet appears to remain scale free and well characterized by stationary statistics [4]. Similarly, the number of links per substrate in the metabolic networks of organisms appears to increase linearly with substrate number [5], the average number of links per scientist in collaboration networks increases linearly over time [6, 7, 8, 9, 10], and languages appear to evolve via accelerated growth [11].

In the main, the chief focus of these studies has been on locating parameter regimes allowing accelerating networks to maintain scale free statistics and thereby to allow continued unconstrained growth. For example, an early study considered a growing network receiving N^α new links for $\alpha > 0$ when the network size is at N nodes, but restricted analysis to the case $\alpha \leq 1$ as "Obviously, α cannot exceed 1 (the total number of links has to be smaller than $N^2/2$ since one may forbid multiple links)." and "The density of connections in real networks remains rather low all the time, so one may reasonably assume that α is small." [12]. Equivalent limits were considered in Ref. [13]. In such restricted parameter regimes networks could maintain scale free statistics, though this result carries the implicit but unexamined finding that alternate parameter regimes permit transitions from stationary to nonstationary statistics. This paper builds on these implicit findings.

Accelerating networks are more prevalent and important in society and in biology than is commonly realized—see the survey in Ref. [14]. In fact, any network that requires functional integration and organization (where the activity of any given node is dependent on the state of the network or different subnetworks) is by definition an accelerating network, that is, as the network expands, the proportion of the network devoted to control and regulation expands disproportionately. This in turn means that all such networks, sooner or later, must be limited in their size and complexity, which limitations can only be breached by changing either the physical nature of the control architecture (a state transition) or by reducing the functional integration. In the latter case, where networks are hitting a complexity limit, further growth in network size will likely display structural transitions from randomly connected, to scale free statistics, to densely connected and perhaps finally to fully connected statistics. Should such networks be unable to successfully complete these transitions for any reason, then it is likely that network growth must cease entirely or that either a transition to a nonaccelerating structure is required to permit further growth or novel technologies must appear allowing the continuation of accelerated growth. Exemplar accelerating networks displaying such size limits or structural transitions include (a) all forms of economic markets where the latest price offered by any participant instantly affects all other participants, (b) industrial companies and sectors implementing a Just-In-Time business model where any worker can halt the entire production system, (c) error propagation networks linking an error source with all affected nodes as studied in software analysis and in models of the propagation of diseases, bushfires, cracks, and electricity grid failures, (d) in any dynamical system dependent on relative quantities so changes in one node instantly affects ev-

*Electronic address: m.gagen@imb.uq.edu.au

ery other node such as relative transcription factor binding probabilities or relative evolutionary fitness, (e) in computer hardware and in cluster and grid supercomputer networks, and (f) in organizational networks [14]. In fact, it is well understood that social networks only take on small world statistics when the network is large enough—in small towns everyone one knows everyone else so social networks are accelerating, and social networks make a transition to small world statistics only as individual nodes saturate their connectivity limits [15]. Similar observations can be made about the scale free Internet and World Wide Web—when sufficiently small, these networks were likely accelerating until connectivity capacities saturated forcing a transition to scale free structures to permit further growth [14].

This paper develops an accelerating network model of prokaryotic (single celled) gene regulatory networks to investigate size and complexity limits inherent in the adoption of an accelerating architecture. Because our focus is on structural transitions, we explicitly do not need to restrict the degree of acceleration to low values of $\alpha \approx 0$. Rather, we permit this parameter to take on any value including $\alpha > 1$ and ensure that the network is not saturated by making link formation probabilistic. The resulting novel "probabilistic" accelerating networks grow by adding on average pN^α new links with $\alpha > 0$ and otherwise arbitrary provided the probability of adding a link is suitably constrained $p \ll 1$ so that total link number remains less than of order N^2.

The gene regulatory model presented here is motivated by comparative genomics findings that the total number of regulatory proteins controlling gene expression (links) scales quadratically with the number of genes or operons (nodes) in prokaryotes [16, 17]. This quadratic growth results as the number of links made by a regulator exploiting homology dependent (sequence specific) interactions scales proportionally to the number of randomly drifting promotor sequences or effectively, with gene number [17]. Hence, gene regulatory networks are inherently accelerating—the probable number of links per regulator pN^α increases linearly with node number with $\alpha = 1$, so consequently, the total number of links scales quadratically as $pN^{\alpha+1}$. In small and sparsely connected networks, most links come from different regulators suggesting that regulator number also scales quadratically with gene number, $pN^{\alpha+1}$. Such an accelerating network would be characterized initially by sparse connectivity at low gene numbers and subsequently by denser connectivity at high gene numbers as networks attempt a transition to a densely connected regime. If the evolving networks can successfully make this transition, the evolutionary record will display a transition in network statistics for some critical network size N_c. Conversely, if these networks, optimized by evolution in the sparse regime, are unable to make the transition to the densely connected regime, the evolutionary record would show a strict size limit $N \leq N_c$ at some critical network size. But this is exactly what is observed. All prokaryotic gene numbers and genomes are indeed of restricted size (less than about 10,000 genes with genomes of between 0.5 and 10 megabases [18]), in contrast to the genomes of multicellular eukaryotes (with for humans, about 30,000 genes and a genome of about 3 gigabases [19, 20]). Ref. [17] predicted the size limit $N_c \leq 20,000$ genes as continued genome growth requires the number of new regulators to exceed the number of nonregulatory nodes.

A satisfactory model of prokaryotic gene regulatory networks requires some novel features. As mentioned above, we introduce probabilistic link formation to allow rapid accelerated growth and correspondingly stricter size limits. (A different but related mechanism was introduced in Refs. [21, 22] which considered the effects of stochastic fluctuations in the number of added links with each additional node.) In addition, we employ directed links and partition nodes into two classes where "regulators" can source outbound regulatory links to regulate other nodes (both regulators and non-regulators), while "non-regulators" cannot source outbound links. (Ref. [23] has previously considered networks of distinguishable nodes.) Further, experimental evidence presented below indicates that the distribution of inbound links is compact and exponential while the distribution of outbound links is long-tailed and likely scale-free. As a result, the heads and tails of our directed links are placed according to two distinct distributions. Altogether, these features allow the reproduction of the observed features of prokaryote gene regulatory networks and satisfactorily predicts the maximum prokaryotic gene count.

Our approach reproducing accelerating network statistics for growing prokaryote genomes complements and informs alternate networking approaches seeking to deduce or simulate the regulatory networks of particular organisms from gene perturbation and microarray experiments [24, 25, 26].

In Section II we canvass the available literature to characterize the statistics of prokaryote gene regulatory networks. This then allows the construction of accelerating growth network models in Section III where we use the continuous approximation and simulations to analyze network statistics. The size constraints inherent in accelerating prokaryote regulatory networks are modelled in Section IV.

II. OVERVIEW OF PROKARYOTE GENE NETWORKS

Ongoing genome projects are now providing sufficient data to usefully constrain analysis of the gene regulatory networks of the simpler organisms. Ref. [16] first noted the essentially quadratic growth in the class of transcriptional regulators (R) with the number of genes (N_g) in

bacteria with the observed results

$$R \propto \begin{cases} N_g^{1.87\pm0.13}, & \text{transcriptional regulation} \\ N_g^{2.07\pm0.21}, & \text{two component systems} \\ N_g^{2.03\pm0.13}, & \text{transcriptional regulation} \\ N_g^{2.16\pm0.26}, & \text{transcriptional regulation.} \end{cases} \quad (1)$$

Here, the top two lines refer to different classes of regulators while the bottom two lines are the results of a cross-checking analysis of two alternate databases. Quoted intervals reflect 99% confidence limits [16]. The explanation for this quadratic growth was that each additional transcription factor doubles the number of available dynamical states which, it was posited, allows for a doubling in the fixation probabilities for this class of genes.

As noted above, Ref. [17] provides an alternate theoretical analysis predicting quadratic growth in any regulatory network exploiting homology dependent interactions and analyzed 89 bacterial and archeael genomes to determine the relations

$$R = \begin{cases} aN_g^b = (1.6 \pm 0.8)10^{-5} N_g^{1.96\pm0.15} & (r^2 = 0.88) \\ pN_g^2 = (1.10 \pm 0.06)10^{-5} N_g^2 & (r^2 = 0.87) \\ cN_g = (0.055 \pm 0.004) N_g & (r^2 = 0.75). \end{cases} \quad (2)$$

In this paper, accelerating networks will be based on the quadratic second line (while nonaccelerating models presented in later work will work with the linear third line [27]). In all cases, the limits reflect 95% confidence levels. For completeness, the data is shown in Fig. 1. The observed quadratic growth implies an ever growing regulatory overhead so there will eventually come a point where continued genome growth requires the number of new regulators to exceed the number of nonregulatory nodes, and based on this, Ref. [17] predicted an upper size limit of about 20,000 genes, within a factor of two of the observed ceiling.

Earlier surveys of bacterial genomes noted that larger genomes harboured more transcription factors per gene than smaller ones [28], with this trend attributed to the need in larger genomes for a more complex network of regulatory proteins to achieve coordinated expression of a larger set of cellular functions, and to selection in complex environments leading to enrichment in transcription factors allowing regulation of gene expression and signal integration. A similar upward trend in the proportion of regulators as a fraction of genome size with increasing genome size was observed in Ref. [29] attributed to a need for an increasing responsiveness in diverse environments, with confirming observations in Ref. [30].

Prokaryotes typically group their DNA encoded genes in operons, co-regulated functional modules of average size 1.70 genes each in *E. coli* which value we treat as typical though in reality, operon size decreases slightly

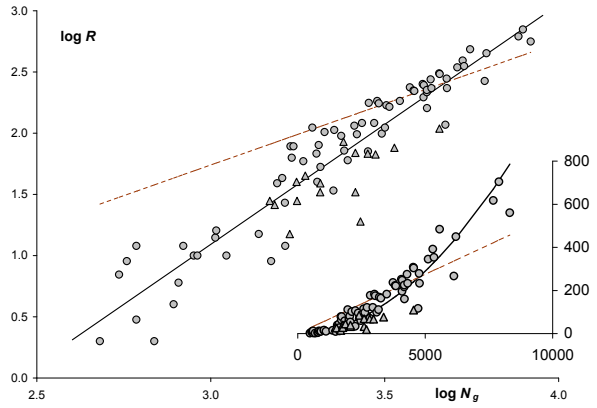

FIG. 1: *Double-logarithmic plot of regulatory protein number (R) against total gene number (N_g) for bacteria (circles) and archaea (triangles), adapted from Ref. [17]. The log-log distribution is well described by a straight line with slope* 1.96 ± 0.15 *($r^2 = 0.88$, 95% confidence interval indicated), corresponding to a quadratic relationship between regulator number and genome size. The inset shows the same data before log-transformation [17]. Dashed lines show the best linear fit to the data* $R = (0.055 \pm 0.004)N_g$ *($r^2 = 0.75$).*

with genome size [31]. Each operon can be either unregulated and so constitutively or stochastically expressed or subject to combinatoric regulation by multiple regulatory protein transcription factors binding to each operon's promotor sequence.

Again assuming that *E. coli* is typical, any given regulatory protein affects an average of about 5 operons with this distribution being long tailed [32] so the majority of regulators affect only one operon though some regulators (CRP) can affect up to 71 operons or 133 genes [33]. (This latter reference estimated that each regulator controls on average 3 genes.) More recent estimates have the transcription factor CRP, a global sensor of food levels in the environment, regulating up to 197 genes directly and a further 113 genes indirectly via 18 other transcription factors [34]. (To observe the long tailed distribution, see Fig. 2 of Ref. [33] and Fig. 4 of Ref. [34].)

However, the number of inputs taken by an operon is characterized by a compact exponential distribution with a rapidly decaying tail so the majority of regulated operons are controlled by a single regulator while very few regulated operons are controlled by four, five, six or seven regulators [32, 33, 34]. In particular, Ref. [33] examined 500 regulatory links from about 100 regulators to almost 300 operons to estimate that each regulated operon takes on average 2 inputs though Fig. 2 of this reference suggests an average input number of about 1.5. Similarly, Ref. [32] suggests that 424 regulated operons receive 577 links giving an average input number of 1.4, while Ref. [34] estimates that 327 regulated operons receive 524 links giving an average input number of 1.6.

III. ACCELERATING PROKARYOTE NETWORK MODELS

We extend the gene network model of Ref. [33] to construct an accelerating network model of prokaryote regulatory gene networks. Prokaryotes typically pack their N_g genes into a lesser number of $N = N_g/g_o$ co-regulated operons where we assume that operons contain exactly $g_o = 1.70$ genes. Of the existing operons, O_r are regulated operons and $O_u = N - O_r$ are unregulated operons. Of the total number of operons, there are R regulatory operons whose regulatory interactions are directed links from regulatory operons to regulated operons. Under the assumption that there is only one regulatory gene per regulatory operon, the observed quadratic relation of Eq. 2 becomes

$$R = pg_o^2 N^2. \tag{3}$$

When regulators and regulatory links are very rare, i.e. when genomes are small, it is likely that every new link is associated with a new regulator so the number of links varies roughly quadratically with operon number. We write

$$L = lN^2, \tag{4}$$

where l denotes the probability of forming a particular beneficial link per operon. The value for l will be approximately pg_o^2, but the exact relation must be derived from the details of the implemented model.

Each regulatory link between nodes is directed, and characterized by two distinct distributions describing respectively the placement of the heads and tails of each link. Only a relatively few nodes are regulatory, and of these, the number of outbound link tails per regulatory node are described by a size dependent long-tailed distribution with average about $\langle t \rangle \approx 5$. Such a long-tailed distribution requires that link tails be preferentially attached to an existing regulatory operon or equivalently, the associated regulated operon must possess one promotor binding site (among others) that binds that particular regulator. Consequently, the preferential selection of regulators means that the promotor sequences of newly regulated nodes cannot be randomly chosen—randomly drifting promotor sequences would be as likely to match any one regulator as another. A plausible physical explanation for the preferential attachment of link tails to existing regulators is that newly fixated operons come largely from gene duplication events [35] where some of the duplicated promotor binding sites are under strong selective constraint while other binding sites and the operon genes can drift freely. Gene duplication then implies that in a genome of size N operons, if some regulator n_j has t_{jN} outbound regulatory links to approximately t_{jN} regulated operons, then the probability that a newly fixated operon is also regulated by n_j is simply the proportion of such regulated operons in the genome, or t_{jN}/N. This implements the required preferential attachment as the resulting rate of growth in the number of links attached to node n_j is also then proportional to t_{jN}. If there is also some probability of the appearance of novel promotor sequences, these combined processes suffice to produce the observed scale free distributions. This model is roughly consistent with recent estimates of the relative contributions to prokaryote genome growth which suggest that horizontal gene transfer rates γ_h are roughly one third of gene loss rates $\gamma_h = \gamma_l/3$ and roughly one half of vertical inheritance or gene genesis rates $\gamma_h = \gamma_v/2$ leading to roughly constant sized genomes over long times (as $\dot{N} \approx \gamma_h + \gamma_v - \gamma_l \approx 0$), while "it is remarkable that phylogenetic distributions of at least 60% [and up to 75%] of protein families can be explained merely by vertical inheritance." [36]. Similarly, three quarters of examined transcription factors in Ref. [34] were two-domain proteins with shared domain architectures leading to the estimate that about 75% of transcription factors have arisen as a consequence of gene duplication (though the joint duplication of regulatory regions and of regulated genes or of transcription factors together with regulated genes is more rare). A further implication of these gene duplication processes is that, in the main, regulators can only appear on entry to the genome—a potential regulator lacking any target matches in a given genome will never form any links when most operons arise from promotor preserving duplication events. This allows us to considerably simplify our model, and hereafter, we only allow regulators to appear on their entry to the genome. Of course, more realistic but considerably more complicated models are possible.

In contrast to the relatively small number of regulatory nodes, all nodes can themselves be regulated by inbound links and in fact, can be multiply regulated as promotor regions can contain more than one binding site. Further, the many used and unused promotor region binding sites broadly sample the space of possible binding sites so only a small fraction of nodes will be regulated by any one regulator. As a result, the number of inbound link heads per node is described by a size dependent exponential distribution with a low average of $\langle h \rangle \approx 1.5$ as typically results from the random or non-preferential attachment of inbound links to operon promotor sequences.

We suppose that the operon network grows by the sequential addition of numbered nodes n_k for $1 \leq k \leq N$, and that at network size k, node n_i $(1 \leq i \leq k)$ has t_{ik} outbound tails and h_{ik} inbound heads. We do not model the many trials of potential genes over many generations and merely include fixated genes in our count—that is, drifting sequence is not counted as part of the fixated genome. This further implies the sequence of established nodes is under severe selective constraint and unable to drift so consequently new links cannot be added between existing nodes. (If a proportion fN of existing nodes can explore novel sequence space in time dt, then the number of new regulators increases as $dR \propto fN^2 dt$, and as N is itself a function of time, this integrates to generate a non-quadratic relation between regulator and operon

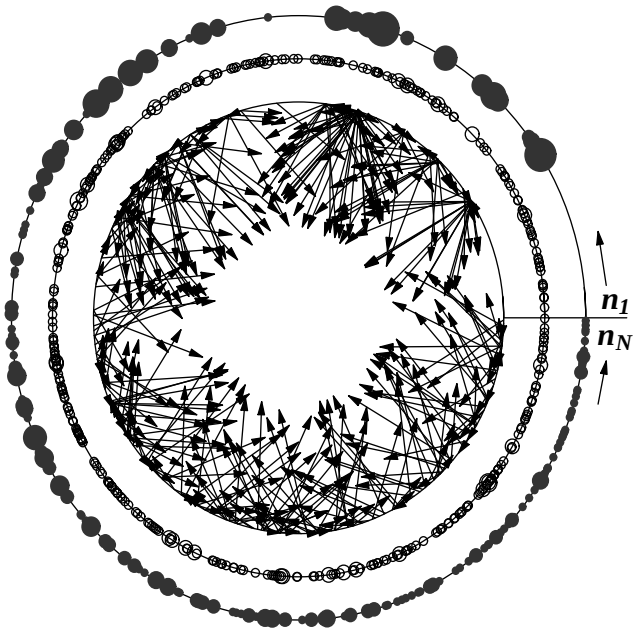

FIG. 2: *An example statistically generated E. coli genome using the later results of this paper where (for convenience only) operon nodes numbered $n_1, \ldots, n_N$ are placed sequentially counterclockwise on a circle in their historical order of entry into the genome. The filled points on the outer circle locate regulators and have radius indicating the number of outbound regulatory links. The open points on the middle circle locate regulated operons and have radius indicating the number of inbound regulatory inputs. The arrows in the inner circle show all directed regulatory links.*

number which is not observed.)

For clarity, Fig. 2 preempts later calculations and depicts a statistically generated version of an *E. coli* genome where nodes are placed sequentially counterclockwise in a circle (for convenience only). Alternative genome models may be distinguished by the age distribution of regulators, regulated operons and their link numbers, and these are indicated in this figure. In particular, Fig. 2 shows a highly nonuniform distribution of regulators and outbound link numbers with gene age in contrast to a uniform distribution of regulated operons and of inbound link numbers. (It will turn out that these latter age-independent distribution are only present when regulator number grows quadratically with genome size.)

These distributions result from the physical processes underlying the formation of regulatory links in prokaryotes. As discussed above, a substantial proportion of the gene regulation network of prokaryotes is enacted via homology dependent interactions as when sequence specified protein transcription factors bind to specific promoter sequences. The undirected nature of evolutionary searches means that gene regulatory networks fundamentally exploit the same sequence matching algorithms used in comparative genetics where the probability of obtaining matches between a single given trial sequence of some small fixed length and an entire genome scales proportionately to genome length—doubling genome length doubles the probability of a match. Hence, the expected number of links formed per regulator scales linearly with present genome size. As the number of source trial sequences also scales with genome length, the expected number of matches between all regulators and all regulated operons scales quadratically with genome length, or effectively, with operon number assuming constant sized operons over the evolutionary record.

As a consequence, on entry into the genome, each new gene has some probability of being a regulator dependent firstly on its suitability to bind DNA and secondly on the linearly increasing expected number of acceptable binding targets present in the genome on entry (or at later times). As discussed above, the predominance of vertical gene genesis events allows a simplified model wherein the probability of a new node being regulatory is determined solely by the number of available links present at the time of entry. We assume then that on entry into the genome each new node n_k can form $2k-1$ links with nodes $n_1, \ldots, n_k$ consisting of a single self-regulatory link from node n_k to itself with probability l, $(k-1)$ regulatory outbound links to the existing nodes each with equal probability l, and, provided that sufficient regulators already exist, $l(k-1)$ inbound regulatory links from some subset of the existing regulators chosen according to preferential attachment. (For consistency, we can only add $\approx lk$ distinct regulatory links to node n_k provided there are at least this many regulators in existence. From Eq. 3, the average number of regulators $pg_o^2k^2$ must be greater than the number of regulatory links lk, and this will be satisfied for $k > l/(pg_o^2) \approx 1$.) As a result, the total number of heads or tails attached to node n_k on entry to the genome ranges between 0 and k, with each link formed with probability l. Hence, the respective probabilities that the initial number of heads $h_{kk} = j$ or the initial number of tails $t_{kk} = j$ for node n_k is

$$P(j,k) = \binom{k}{j} l^j (1-l)^{k-j}, \tag{5}$$

with the proviso that all the inbound links can only be added to node n_k if there is a sufficient number of regulators among the nodes $n_1, \ldots, n_k$. The average number of inbound and outbound links is identical, $\langle t_{kk} \rangle = \langle h_{kk} \rangle = lk$ showing linear growth in link number with increasing network size. The addition of node n_k and its links will increase the probable number of heads attached to earlier nodes n_j for $1 \leq j \leq (k-1)$ so $h_{jk} \geq h_{jj}$, while the probable number of tails outbound from node n_j increases $t_{jk} \geq t_{jj}$ if and only if that node is regulatory with $t_{jj} > 0$.

As regulators can only be created on entry to the genome, the distribution of regulators at any time is specified by the distribution $P(j,k)$ for t_{kk}. Using Eq. 5, the probability that node n_k is a regulator is $1 - P(0,k)$, so for a network of N nodes, the predicted total number of

regulators is

$$\begin{aligned} R &= \sum_{k=1}^{N} \left[1-(1-l)^k\right] \\ &= N - \frac{1-l-(1-l)^{N+1}}{l} \\ &\approx \frac{l}{2}N(N+1). \end{aligned} \tag{6}$$

The exact top line shows the expected behaviour for the number of regulators in the respective limits $l \to 0$ giving $R \to 0$, and $l \to 1$ giving $R \to N$. The approximate relation in the third line can be compared to the observed Eq. 3 and immediately suggests $l \approx 2pg_o^2$, while a fit to the more accurate top line gives the connection probability as

$$l = 1.15 \times 2pg_o^2 \;=\; 7.31 \times 10^{-5}. \tag{7}$$

This probability value suggests an average promotor binding site length of $-\log_4 l = 6.9$ bases. The average number of links per regulator using the second line of Eq. 6 is then approximately $L/R \approx 2$, while the more accurate top line with $N = 2528$ operons for *E. coli* [31] gives $L/R = 2.12$, about a factor of two from the observed value of 5 for *E. coli* [32].

A. Random distribution of regulated operons

The distribution of link heads for all nodes (with possession of a link head designating a regulated node), can be straightforwardly calculated under the assumption that the $t_{kk} \approx lk$ new tails added with node n_k are randomly distributed across the k existing nodes so on average, each existing node receives l links. To build insight, it is useful to consider the general case where $t_{kk} \approx h_{kk} \approx lk^\alpha$ for $\alpha \geq 0$. Setting $\alpha = 0$ adds with some probability a constant number of links with each new node, $\alpha = 1$ adds a linearly growing number of probable links with each new node, $\alpha = 2$ adds a quadratically growing number of probable links with each new node, and so on. The total number of links present in the network is then

$$\int_0^N 2lk^\alpha = \frac{2lN^{\alpha+1}}{\alpha+1} \tag{8}$$

The continuous approximation [37, 38, 39] for links randomly distributed over k existing nodes determines the number of inbound head links for node n_j according to

$$\begin{aligned} \frac{\partial h_{jk}}{\partial k} &= \frac{t_{kk}}{k} \\ &= lk^{\alpha-1}. \end{aligned} \tag{9}$$

This can be integrated with initial conditions $h_{jj} \approx lj^\alpha$ at time j and final conditions $t_{jN} \approx lN^\alpha$ at time N to give

$$h_{jN} = \begin{cases} l + l\ln\frac{N}{j} & \text{if } \alpha = 0 \\ \\ \frac{l}{\alpha}N^\alpha + \frac{l(\alpha-1)}{\alpha}j^\alpha & \text{if } \alpha > 0. \end{cases} \tag{10}$$

Integration of these link numbers over all node numbers j gives the required total number of links as in Eq. 8. For $0 \leq \alpha < 1$, the number of links per node is monotonically decreasing with node number. However, for $\alpha = 1$ and only in this case, the final distribution is independent of node number j because earlier nodes receive exactly enough links from latter nodes to balance the initially biased distribution of heads $h_{jj} \approx lj$, so in the end, all nodes receive on average the same number of inbound regulatory links $\langle h_{jN}\rangle = lN$ for $1 \leq j \leq N$. For faster acceleration rates, $\alpha > 1$, the number of links per node is monotonically increasing as later nodes receive a greater number of links on entry to the genome and this imbalance is not corrected.

The possibility of monotonically increasing numbers of links with node number in accelerating networks has not previously been considered. This possibility requires modifying the usual continuum approach [37, 38, 39] so the final inbound link distribution is obtained via

$$\begin{aligned} H(k,N) &= \frac{1}{N}\int_0^N dj\; \delta(k-h_{jN}) \\ &= \pm\frac{1}{N}\left(\frac{\partial h_{jN}}{\partial j}\right)^{-1} \text{at } [j=j(k,N)], \end{aligned} \tag{11}$$

where $j(k,N)$ is the solution of the equation $k = h_{jN}$. The top line is used when all nodes possess the same average link number while the second line is applicable with the plus (negative) sign when the average numbers of links per node is monotonically increasing (decreasing) with node number. Non-monotonic cases require alternate approaches.

Under quadratic growth in total link number when $\alpha = 1$, and only in this case, the final distribution of link heads is independent of node number and evaluated using Eq. 11 to give

$$\begin{aligned} H(k,N) &= \frac{1}{N}\int_0^N dj\; \delta(k-lN) \\ &= \delta(k-lN). \end{aligned} \tag{12}$$

As expected, a compact final link distribution results when all nodes have an average of $t_{jN} = lN$ inbound regulatory links at time N. This distribution calculated under the continuous approximation equates to one where in reality, each node receives a controlling head with probability l from every other node (though in practise, the total number of received links is of order unity). Hence, for any node in a network of size N, the actual probability of having k heads is

$$H(k,N) = \binom{N}{k} l^k (1-l)^{N-k}. \tag{13}$$

A network simulation with linear growth of link numbers per node model (Eq. 5) serves to validate this predicted final distribution. Fig. 3 compares the predicted distribution $H(k,N)$ against observed distributions for typical simulated networks of various sizes with negligible discrepancies.

For a network of size N, the probability that any given operon is unregulated is $H(0,N)$ so the expected number of unregulated operons summed over all N nodes is

$$O_u = N(1-l)^N. \tag{14}$$

This determines the number of regulated operons as

$$O_r = N - O_u = N\left[1-(1-l)^N\right], \tag{15}$$

showing the expected behaviour as $l \to 1$ giving $O_r \to N$ and $l \to 0$ giving $O_r \approx lN^2 = L$ as each of the sparsely distributed links hits a distinct regulated operon. We note that random gene duplication and deletion events will not change the $H(k,N)$ distribution (other than changing N) as all nodes are identically connected on average. The $H(k,N)$ distribution appears in Fig. 2 which shows a uniform (age-independent) distribution of regulated nodes over the genome, and this uniformity is only expected for $\alpha = 1$ corresponding to linear growth in link numbers per node and quadratic growth in regulator numbers.

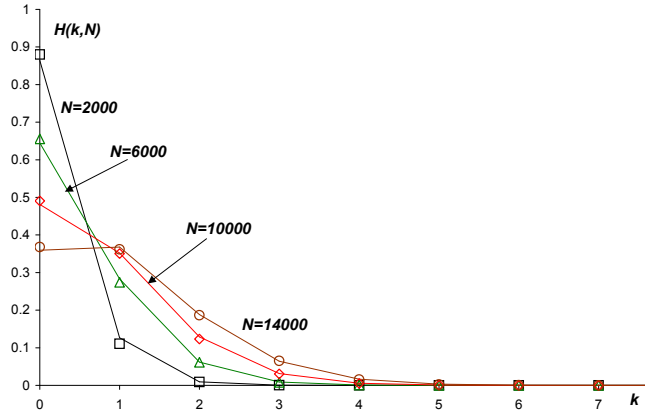

FIG. 3: *A comparison of the predicted distribution of inbound link numbers per node $H(k,N)$ (solid lines) against that observed in simulated networks of various sizes (indicated points) with quadratic growth in the total probable number of randomly attached links.*

These predictions can be compared to observation. For the *E. coli* network of size $N = 2528$ operons or 4289 genes [31], the predicted proportion of regulated operons receiving $k > 0$ inputs is

$$P_h(k) = \frac{H(k,N)}{1-H(0,N)}, \tag{16}$$

and is shown in Fig. 4. Here, the calculated distribution closely approximates the compact exponential distribution observed for *E. coli* shown in Fig. 2(d) of Ref. [31]

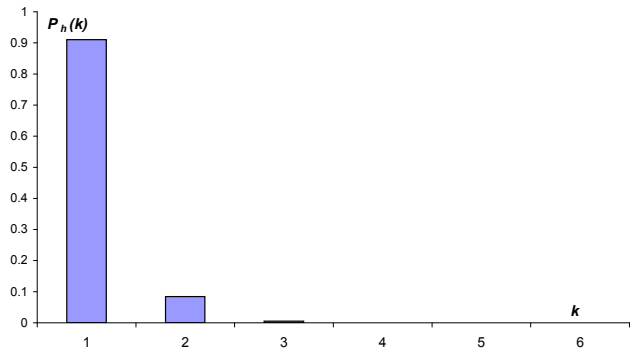

FIG. 4: *The predicted proportions $P_h(k)$ of the regulated operons of E. coli taking multiple regulatory inputs for a genome of $N = 2528$ operons. This distribution closely approximates that observed for E. coli in Fig. 2(d) of Ref. [31] and of Fig. 5 of Ref. [34].*

and of Fig. 5 of Ref. [34], though it underestimates the numbers of regulated operons with 4, 5, 6 and 7 inputs—essentially no regulators are predicted to have 5 or more inputs for genomes of size $N = 2528$ operons. In addition, the average number of inbound regulatory links per operon (for all operons) is $\langle k \rangle = L/N = lN = 0.19$, while the average number of inbound regulatory links for regulated operons is $\langle k_r \rangle = L/O_r \approx 1$. A more accurate calculation using the specific values for *E. coli* gives $\langle k_r \rangle = L/O_r = 1.10$, very close to the *E. coli* value of 1.5 or 1.6 noted in Refs. [32, 33, 34].

B. Scale-free distribution of regulator operons

On entry into the genome, node n_k sources on average lk outbound regulatory links and this linear growth in link number means that more recent nodes are more likely to be immediately regulatory and more likely to be highly connected on genome entry. However, node n_k will also receive on average lk inbound regulatory links whose tails will be preferential attached to existing regulators. The final distribution of link number with age will depend on the rate at which earlier nodes under preferential attachment can attract links relative to the linearly increasing link numbers of later regulatory nodes.

On entry at time k, node n_k receives $h_{kk} \approx lk$ inbound links from existing regulatory nodes in the set $n_1, \ldots, n_k$. As previously, we gain insight by considering the general case where $t_{kk} \approx h_{kk} \approx lk^\alpha$ for $\alpha \geq 0$ (though we continue to use the distribution $P(j,k)$ of Eq. 5 to determine both the number of links j prior to exponentiation and regulatory probability so consequently the number of regulators continues to increase quadratically according to Eq. 6). As a result, the need to ensure that all regulatory links to node n_k are distinct requires that new link number lk^α be less than the number of existing regulators lk^2 requires $\alpha \leq 2$. The h_{kk} new tails added with node

n_k are preferentially attached to the existing regulatory nodes n_j with probability proportional to the number of existing regulatory links for that node at time k, i.e. t_{jk}. Using the continuous approximation [37, 38, 39], the rate of growth in outbound link number for node n_j is then approximately

$$\frac{\partial t_{jk}}{\partial k} = h_{kk}\frac{t_{jk}}{\int_0^k t_{jk}\, dj}. \tag{17}$$

The denominator here is a probability weighting to ensure normalization and is the total number of outbound links for all nodes. Following [1], we can evaluate the denominator using the identity

$$\frac{\partial}{\partial k}\int_0^k t_{jk}\, dj = \int_0^k \frac{\partial}{\partial k} t_{jk}\, dj + t_{kk}. \tag{18}$$

This can be evaluated using Eq. 17 noting $t_{kk} \approx h_{kk} \approx lk^\alpha$ giving

$$\frac{\partial}{\partial k}\int_0^k t_{jk}\, dj = 2lk^\alpha, \tag{19}$$

which can be integrated determining the denominator of Eq. 17 to be

$$\int_0^k t_{jk}\, dj = \frac{2l}{\alpha+1}k^{\alpha+1}. \tag{20}$$

This is in agreement with Eq. 8. Substituting this value into Eq. 17 gives

$$\frac{\partial t_{jk}}{\partial k} = \frac{\alpha+1}{2}\frac{t_{jk}}{k}. \tag{21}$$

Finally, this can be integrated with initial conditions $t_{jj} \approx lj^\alpha$ at time j and final conditions t_{jN} at time N to give

$$t_{jN} = lN^{\frac{\alpha+1}{2}} j^{\frac{\alpha-1}{2}}. \tag{22}$$

Again we find that the respective choices $\alpha < 1$ and $\alpha > 1$ lead to monotonically decreasing and increasing numbers of links per node as a function of node number, while setting $\alpha = 1$ ensures the number of links per node is independent of node number. In this case, the preferential attachment of links to earlier nodes does indeed act to cancel the initial bias in link number towards later nodes. It is also apparent that when $\alpha = 1$, the limit $l \to 1$ implies all nodes possess exactly N links as expected for a fully connected regular network. (Preferential attachment cannot distort connectivity numbers in this case as all nodes have an equal number of links.) Additionally, in the limit $l \to 0$ we have $t_{jN} = 0$ as required for an entirely disconnected network. The case $\alpha = 0$ duplicates results found for growing networks which add a constant number of links with each new node subject to preferential attachment [2].

As previously, it is straightforward to calculate the final outbound link distribution in the case $\alpha = 1$ using Eq. 11. This gives

$$\begin{aligned} T(k,N) &= \frac{1}{N}\int_0^N dj\ \delta(k - lN) \\ &= \delta(k - lN). \end{aligned} \tag{23}$$

Again, we find the expected compact distribution resulting when all nodes possess the same average number of links. This raises the question however, of how it is that a probabilistic accelerating network subject to preferential attachment can end up with all nodes possessing the same average number of links? The answer lies in our use of two classes of distinguishable nodes, regulators and non-regulators, which requires that we take into account the known distribution of regulators with node number over the genome. The average link number per node at node n_j (Eq. 22) equates to the product of the average number of link tails per regulator at node n_j, denoted $t_r(j,N)$, and the average number of regulators per node at node n_j, denoted $\rho(j)$. This latter density is $\rho(j) = dR(j)/dj \approx lj$ by Eq. 6, so by definition, we have

$$t_{jN} = t_r(j,N)\rho(j), \tag{24}$$

giving

$$t_r(j,N) = N^{\frac{\alpha+1}{2}} j^{\frac{\alpha-3}{2}}. \tag{25}$$

Hence, for $\alpha < 3$, the average number of links per regulator is a decreasing function of node number j as the growing number of links added to recent nodes is insufficient to outweigh the effects of preferential attachment which more rapidly increases the number of links attached to early nodes. In particular, for $\alpha = 1$ with the addition of a linearly increasing number of links per node, the average number of regulatory links per regulator scales inversely with node number j. In other words, the density of regulators is very low at small node numbers j while the very few regulatory nodes in this stretch of the genome are heavily connected due to preferential attachment so as to maintain the constant average of Eq. 22. (See Fig. 2.)

The $t_r(j,N)$ distribution contains information about both node connectivity and node age and so approximates genome statistics (simulated or observed) when all of this information is available. However, it is usually the case that node age information is unavailable necessitating calculation of connectivity distributions that are not conditioned on node age. This effectively requires binning together all nodes irrespective of their age to obtain a final link distribution. In the case of linearly growing number of links per node, $\alpha = 1$, the delta function of Eq. 11 is resolved by the equality $j = N/k$ giving the final distribution as

$$\begin{aligned} T(k,N) &= -\frac{1}{N}\left(\frac{\partial j}{\partial k}\right) \\ &= \frac{1}{k^2}, \end{aligned} \tag{26}$$

which, as required, is normalized as $\int_1^\infty T(k,N) = 1$. The expected proportion of regulators $P_t(k)$ possessing k links is then obtained by integrating the continuous distribution of Eq. 26 over appropriate ranges $[1, 3/2]$ or $[k-1/2, k+1/2]$ to obtain

$$P_t(k) = \begin{cases} \frac{1}{3} & k = 1 \\ \\ \frac{4}{4k^2-1} & k > 1. \end{cases} \quad (27)$$

These theoretical predictions compare well to simulations of networks of various sizes with linearly increasing numbers of probable links per node and subject to preferential attachment. Fig. 5 shows simulated outbound link distributions which are long-tailed and scale free with probabilities scaling roughly as $P_t(k) \propto k^{-2}$ for large k. The $P_t(k)$ distribution shows a full one third of regulators have only one link, while 60% have two or fewer links, and 71% have three or fewer links. Fig. 6 shows the long-tailed distribution $P_t(k)$ expected for a simulated *E. coli* network of $N = 2528$ operons with preferential attachment of links. This figure shows marked similarities to the long-tailed distribution of *E. coli* shown in Fig. 2(c) of Ref. [31]. In particular, the expected number of regulators with k links is $P_t(k)R(N)$ with the number of regulators $R(N)$ obtained from Eq. 6 (or from observation). For *E. coli*, this predicts the probable existence of about one regulator possessing link numbers in each of the respective ranges between $[40, 49]$ links, between $[50, 64]$ links, between $[65, 94]$ links, between $[95, 169]$ links, and between $[170, 700]$ links for instance. (This approximates the connectivity of the global food sensor CRP which regulates up to 197 genes directly [34].) The average of the $P(k)$ distribution (as well as the $t_r(j,N)$ distribution) is formally undefined as long as the integration limits are taken to infinity. However, in a network of N nodes, a regulator can practically only regulate a total of N nodes, and this cutoff allows us estimate the average connectivity per regulator (complementing previous estimates following Eq. 7). Using the cutoff and approximating the summation via an integral, the average connectivity per regulator in a network of N nodes is

$$\begin{aligned} \langle k \rangle &= \sum_{k=1}^{N} k P_t(K) \\ &= \frac{1}{3} + \frac{1}{2} \ln \left(\frac{4N^2 - 1}{15} \right), \end{aligned} \quad (28)$$

(or simply $\ln N$ using the continuous distribution of Eq. 26.) The average number of links per regulator for *E. coli* from Eq. 28 is $\langle k \rangle = 7.51$ (or 7.83 using the simpler derivation), which again compares well to the observed value of 5 in *E. coli* [32].

IV. INHERENT PROKARYOTE SIZE LIMITS

The accelerating nature of regulatory gene networks necessarily means that these networks must exhibit a transition at some critical network size either to a nonaccelerating architecture permitting continued growth or must cease growth entirely, and we now seek to predict the location of this transition point and compare it to the evolutionary record. We begin by examining an overview of the accelerating genome model. Fig. 7 shows that linear growth in link numbers per node ($\alpha = 1$) allows a quadratic growth in the total number of links (Eq. 4) despite each of the number of regulators (Eq. 6) and the number of regulated nodes (Eq. 15) asymptoting to some

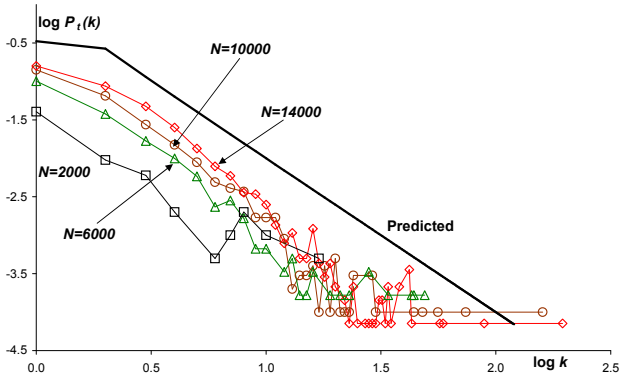

FIG. 5: *A simulation of the proportion of outbound links per regulator $P_t(k)$ in networks of various sizes with linear growth in the probable number of links per node preferentially attached to regulatory nodes. The log-log plot shows slopes of roughly -2 in agreement with theoretical predictions (heavy solid line) of a long-tailed scale free distribution $P_t(k) \propto k^{-2}$.*

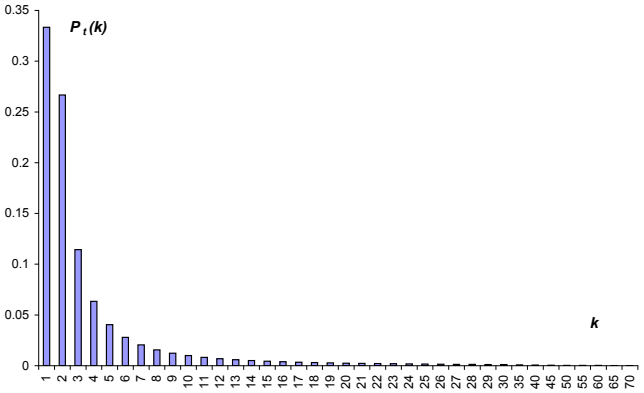

FIG. 6: *The predicted proportion of regulatory operons $P_t(k)$ regulating k different operons for a simulated E. coli genome with $N = 2528$ operons. As expected, most regulators regulate only one other operon, though a small number of regulators can regulate more than 40 operons. This distribution closely approximates the observed proportions for E. coli in Fig. 2(c) of Ref. [31] and Fig. 4 of Ref. [34], and predicts the probable existence of about one E. coli regulator possessing link numbers in each of the respective ranges between $[40, 49]$ links, between $[50, 64]$ links, between $[65, 94]$ links, between $[95, 169]$ links, and between $[170, 700]$ links, and so on.*

fraction of N after an initial period of quadratic growth. For large genomes, almost all new nodes will be regulators and densely connected into the existing network which will then multiply regulate almost every node.

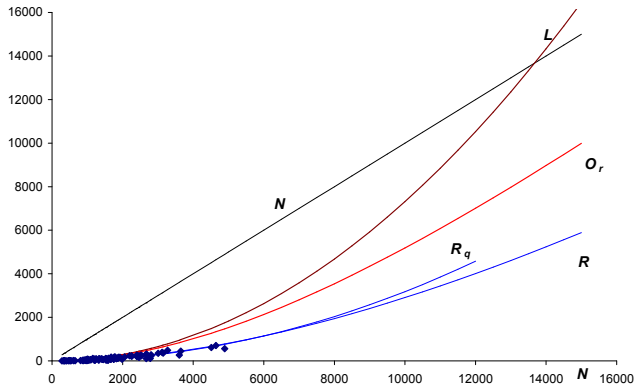

FIG. 7: *The quadratic growth in the number of regulatory links L, and the asymptoting quadratic growth of regulatory operons R and of regulated operons O_r in relation to the total number of operons N. Actual numbers of regulators for 89 prokaryote genomes are shown (solid points), while the non-asymptoting fitted quadratic curve R_q is shown for comparison. The observed maximum size of prokaryote genomes (of order 10,000 genes or about 6,000 operons) lies near the transition point between sparse and dense connectivity as an increasing proportion of operons become linked into the regulatory network.*

The transition from sparse to dense connectivity occurs as an increasing proportion of operons become linked into the regulatory network leading to the emergence of a single giant component of fully connected nodes. One way to highlight this transition is by determining the proportion of transcription factor which control downstream regulators as such linkages create the single giant component. The proportion of regulators controlling regulators is

$$\begin{aligned} P_{rr}(N) &= \frac{1}{R(N)} \sum_{k=1}^{N} \left[1-(1-l)^k\right] \frac{N}{k} \frac{R(N)}{N} \\ &\approx lN. \end{aligned} \tag{29}$$

Here, the first fraction on the RHS normalizes the proportion in terms of the number of regulators $R(N)$ (Eq. 6), the first term in the summation is the probability that node n_k is a regulator, the second term is the average number of regulatory outbound links for this regulatory node $t_r(k,N)$ at network size N (Eq. 25 with $\alpha = 1$), and the third term approximates the probability that these nodes link to one of the existing regulators under random attachment. (If the very first and very last terms are dropped, the remaining summation over all nodes of the probability that n_k is regulatory with the stated number of links equates to the total number of links in the network $L \approx lN^2$. This is the more accurate version of the calculation leading to Eq. 25.) Hence, the proportion of regulators which control transcription factors scales linearly with network size and equals 15% for an $N = 2000$ network, 29% for $N = 4000$, 44% for $N = 6000$, 59% for $N = 8000$, 73% for $N = 10000$, and 88% for $N = 12000$ operons (after which the approximations made break down). Naturally, when most regulators themselves control other regulators, then the entire regulatory network will consist of a single giant component. These ratios compare reasonably well with those observed in *E. coli* where Ref. [34] noted that of 121 transcription factors for which one or more regulatory genes are known, 38 factors or 31.4% regulate other transcription factors. The approximate second line of Eq. 29 with $N = 2528$ for *E. coli* determines this proportion as $P_{rr} = 18.5\%$ while the more accurate top line gives the proportion of regulators which control transcription factors as $P_{rr} = 17.7\%$, giving a reasonable match between prediction and observation.

As the proportion of regulators of transcription factors rises, the probable length of regulatory cascades will increase. In fact, the proportion of regulators taking part in a regulatory cascade of length $n \geq 1$ is

$$p_n = (1 - P_{rr}) P_{rr}^{n-1}. \tag{30}$$

This equation can be obtained from a tree of all binary pathways which at each branching point either terminate with probability $(1 - P_{rr})$ or cascade with probability P_{rr}. As such, the probable cascade length is negligible when the proportion of regulators controlling regulators is small $P_{rr} \ll 1$ but can become large as P_{rr} itself increases. As P_{rr} is indeed large for networks of size $N > 6000$, this again suggests that long cascades of regulatory interactions will lead to the coalescing of a single giant component in this regime. Again, the calculated lengths of regulatory cascades can be compared to those in *E. coli* where the number of cascades of regulated transcription factors observed in a particular set of regulatory interactions was 23 two-level cascades or 37.7%, 32 three-level cascades or 52.5%, and 6 four-level cascades or 9.8% [34]. As one-level or autoregulatory interactions are not included in this observation, the predicted proportions for *E. coli* are $\bar{p}_n = p_n/(1-p_1)$ with $P_{rr} = 17.7\%$ giving 82% two-level cascades, 15% three-level cascades, 3% four-level cascades, 1% five-level cascades, and so on. It is seen that the theoretical predictions overestimate the proportion of two-level cascades and underestimate the number of three-level and higher cascades probably because of selection pressures not included in the model. Lastly, we note that the number of cycles involving closed regulatory loops of size greater than one (i.e. involving more than autoregulation) in the examined portion of the *E. coli* regulatory network is zero reflecting that feedback loops in these organisms are carried out at the post-transcriptional level involving metabolites such as appear in the *lac* operon [32, 33, 34].

We note that our model is entirely unable to explain the high proportion of autoregulation observed in *E. coli* with various estimates that 28.1% [40], 50% [32] and

46.9% [33] of regulators are autoregulatory. The predicted proportion of autoregulators is approximated by replacing the very last fraction (R/N) in Eq. 29 by the term $1/N$ giving the probability that a self-directed link is formed, leading to the expected autoregulatory proportion $\approx 2/N \approx 0.08\%$ for *E. coli*. This failure likely reflects the action of selection processes promoting spatial rearrangements of entire regulons on the genome and the internal shuffling of genes and promotor units. Such reorganizations of duplicated gene regions (presumably shuffling genes and promotor regions) have been common in *E. coli* allowing for instance, spatial regulatory motifs whereby the promotors of colocated (overlapping) and often co-functional operons transcribed in opposing directions can interfere [41].

The transition point from sparse to dense connectivity can be roughly located using the continuum approximation [37, 38, 39]. These methods have not previously been used for this purpose (to our knowledge) and we first validate their use by deriving the known result that non-growing random graphs of N nodes connected by an increasing number of L undirected links undergo a phase transition from sparse to dense connectivity when $L = N/2$ [42]. As the number of links L grows, the N nodes are interlinked into firstly separate islands of size s_i nodes for $i = 1, 2, \ldots$ which eventually link up to form a giant component designated s_1 containing essentially all nodes $s_1 \approx N$. The largest component grows whenever a newly added link has either its head or tail in island s_1 (with probability s_1/N) and the other outside it (with probability $(N - s_1)/N$) leading to a size increment equal to the average size of the external islands ($\langle s_{j\neq 1} \rangle$), giving

$$\frac{ds_1}{dL} = \left[2\frac{s_1}{N}\frac{(N - s_1)}{N} \right] \langle s_{j\neq 1} \rangle. \qquad (31)$$

Numerical or analytic integration of this equation with initial conditions $s_1 = 2$ when $L = 1$ and assuming the average size of external smaller islands is $s_1/2$ shows the largest island saturating the entire network when $L = N/2$ as expected. (This simple approach is indicative only and is quite sensitive to for instance, the assumed average size of external islands.)

This result suggests the following transition point in directed regulatory gene networks. Each undirected (i.e. bidirectional) link in random graph theory is equivalent to two directed links allowing bidirectional traffic between any two nodes, suggesting a transition point in directed graphs at roughly $L = N$. This analysis suggests that the largest component is expected to saturate the entire network when link number $L \approx N$ or $N = 1/l = 13,677$ (see Fig. 7). In turn, this suggests that for $N < 13,677$ a typical network likely consists of isolated trees, while if $N > 13,677$ the network likely consists of a single giant cluster where almost every node is connected to all others via intermediate links. When the link number is very large, $N \gg 13,677$, then the network becomes regularly connected [2]. As prokaryote regulatory networks likely consist of functionally distinct regulated modules [33, 43], it is unlikely that prokaryotic gene networks can successfully operate in the fully connected regime suggesting that prokaryote genome sizes are size constrained $N \leq 13,677$. In fact, the previously noted absence of regulatory cycles in *E. coli* [32, 33, 34] likely reflects the evolutionary importance of maintaining disjoint and non-interfering regulatory units.

These results of random graph theory are suggestive only, and we now turn to consider the size of the largest connected island in prokaryote gene networks featuring directed links whose tails are preferentially attached to regulators and whose heads are randomly distributed over all existing nodes. A further difference is that prokaryote regulatory networks are themselves growing with each added node accompanied by a probabilistic number of links. In addition, we define an island to consist of all nodes which are linked regardless of the orientation of all links and so effectively treat links as being undirected. This is because a regulator can potentially perturb every node downstream to it including those nodes downstream of other regulators and so can modify the regulatory effects of other regulators—essentially, if the downstream effects of different regulators eventually intersect, we count these regulators in the same island. (Other definitions of islands could be used.)

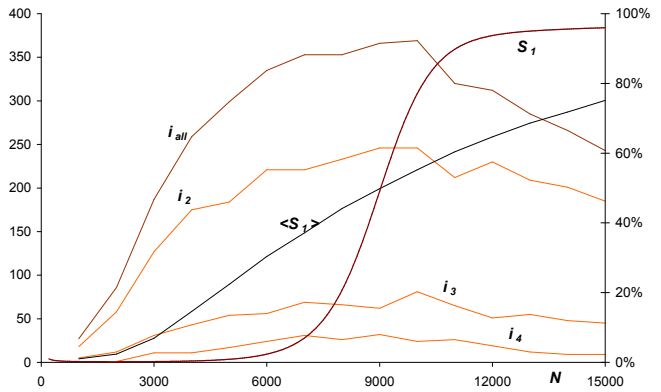

FIG. 8: *The total number of discrete disconnected islands* $i_{\rm all}$*, the number of islands with respectively two* (i_2)*, three* (i_3) *and four* (i_4) *members (left hand axis), and the simulated* ($\langle s_1 \rangle$) *and predicted* (s_1) *size of the largest island measured as a proportion of nodes for various genome sizes (right hand axis). The simulations show the largest island contains* $\langle s_1 \rangle = 50\%$ *of all nodes at a critical network size of* $N_c = 9,029$ *nodes. The input parameters of the predicted curve* s_1 *are set so* $s_1 = \langle s_1 \rangle$ *at this point.*

The dominant (but not sole) mechanism by which island s_1 can grow is for the newly added node n_k to either (a) be a regulator (with probability $[1 - (1 - l)^k]$) and establish an outbound regulatory link to some existing node in s_1 (with probability s_1/k) while at the same time accepting a regulatory link (with probability $[1-(1-l)^k]$) from a node in a different island $s_{j\neq 1}$ (with probability $(k - s_1)/k$), or (b) accept an inbound regulatory link

(with probability $[1-(1-l)^k]$) from a regulator in island s_1 (with probability s_1/k) while establishing a regulatory link (with probability $[1-(1-l)^k]$) to some node in a different island $s_{j\neq 1}$ (with probability $(k-s_1)/k$). (Here, we assume that regulators are uniformly distributed over islands and the number of links within an island scales with the size of the island to crudely model preferential attachment.) The result is that island s_1 grows by the size of the second island assumed to be $s_{j\neq 1}$. Altogether, the rate of growth in the size of island s_1 is then

$$\frac{ds_1}{dk} = 2\left[1-(1-l)^k\right]^2 \frac{s_1[k-s_1]}{k^2}\langle s_{j\neq 1}\rangle. \qquad (32)$$

For initial conditions, we assume that a first link appears when the genome has $(pg_0^2)^{-1/2} = 177$ nodes ($s_1(177) = 2$). Simulations show that sufficient small islands are created to ensure $\langle s_{j\neq 1}\rangle$ remains roughly constant and equal to $\langle s_{j\neq 1}\rangle = 2.72$, though matching the simulated and predicted curves at the 50% point requires setting $\langle s_{j\neq 1}\rangle = 30$. This is reasonable given the approximations made. Fig. 8 shows the size of the largest island s_1 as a proportion of all nodes. A single giant component is expected to form at a critical genome size of $N_c = 9,029$ operons defined as the point where the simulated proportion of nodes in the giant component is 50%. (Choosing a parameter setting of 40% would also be justifiable and would lead to an exact match between predicted and observed maxima.) Unlike random graph theory, this critical point applies to all growing genomes as it is determined by the value of the link formation probability l. Genomes of smaller size than this critical value $N < N_c$ are expected to be sparsely connected so the network consists of multiple discrete connected islands (as in *E. coli* [33]), while genomes of larger size $N > N_c$ are expected to be densely connected into a single giant component where every regulator eventually perturbs the downstream effects of every other regulator.

Simulations of example genomes of various sizes spanning this critical network size confirm the adequacy of the continuum treatment. Fig. 8 shows the number of all discrete islands as well as the number of islands containing two, three and four components. In the vicinity of the critical genome size $N_c = 9,029$, the number of discrete interconnected islands begins to decline as the growing number of links connects more and more islands into the single giant component. The size of the simulated giant component as a proportion of genome size is also shown. This figure suggests that the *E. coli* genome of $N = 2528$ operons should possess a giant component containing about 5% of all nodes (about 100 nodes) which can be compared with the observation that about 70% or 300 operons of the examined regulatory and regulated operons (but not including unregulated and nonregulatory operons) could be loosely grouped into 3-6 "dense overlapping regulons" or DORS while the remaining operons appeared as disjoint systems with most containing 1-3 operons but some containing up to 25 operons [32].

The critical network size of $N_c = 9,029$ operons or about $N_g = 15,349$ genes corresponds to the point where growing regulatory networks exploiting accelerating links can no longer maintain discrete functional units, islands, of interconnected nodes. Larger genomes are densely connected into a single giant component where, eventually, any regulator can perturb the downstream effects of every other node so for instance, it is unlikely that the discrete network motifs found in the *E. coli* regulatory network [32] can survive in this regime. This massive increase in perturbative effects immeasurably increases the difficulty of the evolutionary search process, leading to an expectation that the rate of evolutionary change will drastically slow when growing genome sizes reach criticality $N \approx N_c$. From a biological point of view, it is relatively easy to understand why the critical network size N_c acts as an upper size limit. The accelerating nature of the prokaryote regulation network means that larger networks can add new nodes only be integrating an increasing number of links to gain evolutionary benefits. Of course, the probability of finding lN beneficial links is a rapidly decreasing function of N. It is relatively easy to find a beneficial regulator making only of order one link to existing genes (only billions of trials are needed say), but much harder when the regulator is making an average of five links with existing genes (many trillions of trails are needed). Essentially, the more links that must be beneficially integrated, the longer the evolutionary search task and the slower the rate of evolutionary change.

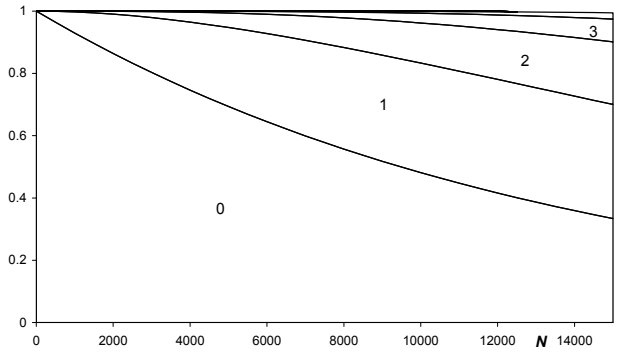

FIG. 9: *The predicted proportion $P_o(k,N)$ of operons with $0,1,2,\ldots$ regulatory inputs as a function of network size. Small networks mainly possess unregulated operons, while networks of large size have a significantly reduced number of unregulated operons with many operons taking large numbers of regulatory inputs.*

Many other statistical measures suggest that the regulatory mechanisms optimized to perform in a sparsely connected network will not necessarily operate in a densely connected network—evolution cannot foresee later needs. In particular, the proportion of operons n_j which are regulated by k inputs is, using Eq. 13, given

by

$$P_o(k,N) = \frac{1}{N}\sum_{j=1}^{N} H(k,N) \;=\; \binom{N}{k} l^k (1-l)^{N-k}. \quad (33)$$

This distribution increases with increasing network size and is shown in Fig. 9 making it clear that small networks mainly possess operons which are either entirely unregulated or regulated by only one or a few regulators. In contrast, large networks ($N > N_c$) have only a small proportion of operons which are unregulated while the majority of operons take between one or more regulatory inputs. It is a more difficult evolutionary task to integrate many inputs to achieve a beneficially regulated output again suggesting that prokaryote regulatory networks featuring accelerating growth in link number are size limited due to their regulatory architecture.

Another way to suggest the strict size limits imposed by the accelerating growth of regulatory links is to consider the probability that the most recently added node n_N in a network of size N immediately becomes regulatory. Using Eq. 5, node n_N is a regulator with probability

$$P_r(N) = \sum_{i=1}^{N} P(i,N) \;=\; 1-(1-l)^N. \quad (34)$$

This probability tends to unity as network size increases, and in particular, surpasses about 50% when networks consist of N_c operons—see Fig. 10. At about this stage, large networks cannot add a new node without it having a significant probability of modifying the dynamics of existing nodes. This immeasurably increases the difficulty of the evolutionary task and again suggests a maximum size limit to prokaryote gene regulatory networks.

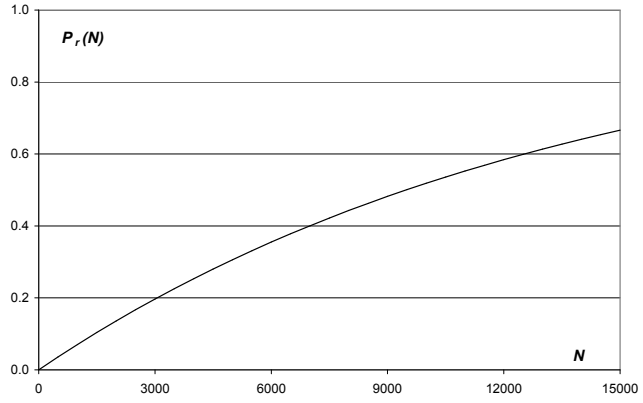

FIG. 10: *The rapidly increasing probability $P_r(N)$ that the most recently added node n_N in a network of size N nodes is immediately regulatory on its appearance in the genome. For network sizes greater than about $N_c = 9,029$ operons, the probability that all new nodes are immediately regulatory exceeds about 50%.*

If the accelerating regulatory networks of prokaryotes were able to operate in the densely connected regime, the evolutionary record might be expected to show prokaryotes of arbitrarily large genome size with a transition in connectivity statistics at some critical genome size of about $N_c \approx 9,029$. Conversely, should these regulatory networks be unable to operate in the densely connected regime, then the evolutionary record should show a maximum size limit to prokaryote genome sizes of about $N_c \approx 9,029$ operons or about $N_g = 15,349$ genes, close to the observed upper limit.

V. CONCLUSION

In this paper, we generalize models of accelerating networks by including probabilistic links to allow arbitrarily rapid acceleration rates leading to structural transitions in growing networks sometimes severe enough to strictly constrain network size. These structural transitions from sparse to dense connectivity are made more difficult by any additional steric or logical limitations on combinatoric control at any given promotor. Such transitions are in sharp contrast to the stationary statistics and unbounded growth potential of non-accelerating scale free and exponential networks. These probabilistic accelerating networks were applied to model prokaryote regulatory networks which exploit a quadratic growth in the number of regulators and regulatory links with genome size as established via comparative genomics programs. Our models predict a maximum genome size of $N_c \approx 9,029$ operons or about $N_g = 15,349$ genes for prokaryotes, closely approximating the observed maximum. We further validated our model by making a detailed comparison of predicted and observed results for *E. coli*, and achieved satisfactory matches for respectively, the number of observed regulators, an average promotor binding site length of about 7, the long tailed distribution of outgoing regulatory links with an average of between 2.12 and 7.51 (compared to 5), the exponential distribution of incoming regulatory links with an average of around 1.10 (compared to 1.5), the proportion of regulators controlling regulators of around 17.7% (compared to 31.4%), and the probable length of regulatory cascades and the absence of regulatory loops. Our approach is unable to explain the high proportion of autoregulation observed in *E. coli* [32] and this failure likely points to selection for genome reorganizations leading to spatial arrangements of operons allowing joint regulation [41] which is not included in this model. Further, this approach does not include selection pressures ensuring that similarly regulated islands or modules share common functionality [32], or other regulatory mechanisms influencing both the transcription and translation of transcription factors including micro-RNAs and other chemical mechanisms and mediators (see for instance [44]).

However, the many successes of the accelerating network model of prokaryote regulatory networks are meaningless if similar results can be achieved via non-accelerating network models. In later work, we will show

that the two simplest non-accelerating network models fail to explain either the observed quadratic growth of regulator number with genome size or the detailed statistics pertaining to the *E. coli* genome [27]. In addition, the simplifying assumption adopted here that gene duplications ensure that operons become regulatory only on entry to the genome will be dropped in later work. This will develop a more realistic model including separate physical processes for transcription factor binding to DNA and for establishing regulatory links with regulated operons where links can form at any time.

This work has wider significance due to the still common presumption in molecular biology that "What was true for *E. coli* would also be true for the elephant" capturing the notion that the mechanisms operating in prokaryotes are exactly identical to those operating in complex multicellular eukaryotes. In this picture, eukaryotes are merely enlarged prokaryotes. The results of this paper indicate that this is not possible—the accelerating nature of regulatory networks necessarily implies that eukaryotes cannot be scaled up prokaryotes and that the (likely) accelerating regulatory networks of eukaryotes must be exploiting novel regulatory mechanisms. The successful modelling of these mechanisms will likely require incorporating computationally complex technologies [45, 46, 47] into an accelerating network model, and this also will be addressed in later work.

[1] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.

[2] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos. Power-law relationships of the Internet topology. In L. Chapin, J. P. G. Sterbenz, G. Parulkar, and J. S. Turner, editors, *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 251–262, New York, 1999. ACM Press.

[4] A. Vásquez. Large-scale properties and dynamical properties of the Internet. *Physical Review E*, 65(5):066130, 2002.

[5] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[6] S. N. Dorogovtsev and J. F. F. Mendes. Scaling behaviour of developing and decaying networks. *Europhysics Letters*, 52(1):33–39, 2000.

[7] A. Vásquez. Knowing a network by walking on it: Emergence of scaling. Eprint cond-mat/0006132, 2000.

[8] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. Eprint cond-mat/0104162, 2001.

[9] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002.

[10] A. Vásquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104, 2003.

[11] S. N. Dorogovtsev and J. F. F. Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London B*, 260:2603–2606, 2001.

[12] S. N. Dorogovtsev and J. F. F. Mendes. Effect of the accelerating growth of communications networks on their structure. *Physical Review E*, 63:025101(R), 2001.

[13] P. Sen. Accelerated growth in outgoing links in evolving networks: Deterministic vs stochastic picture. 2003. arXive:cond-mat/0310513. See http://arxiv.org/abs/cond-mat/0310513.

[14] M. J. Gagen and J. S. Mattick. Inherent limitations of "accelerating" networks in biology and society. *In Preparation*, 2003.

[15] D. J. Watts. Networks, dynamics, and the Small-World phenomenon. *American Journal of Sociology*, 105(2):493–527, 1999.

[16] E. van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genetics*, 19(9):479–484, 2003.

[17] L. J. Croft, M. J. Lercher, M. J. Gagen, and John S. Mattick. Is prokaryotic complexity limited by accelerated growth in regulatory overhead? arXive:q-bio.MN/0311021. See http://arxiv.org/abs/q-bio.MN/0311021, 2003.

[18] S. Casjens. The diverse and dynamic structure of bacterial genomes. *Annual Review of Genetics*, 32:339–377, 1998.

[19] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[20] J. C. Venter, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[21] Z. Liu, Y.-C. Lai, and N. Ye. Statistical properties and attack tolerance of growing networks with algebraic preferential attachment. *Physical Review E*, 66:036112, 2002.

[22] K.-I. Goh, B. Kahng, and D. Kim. Fluctuation-driven dynamics of the Internet topology. *Physical Review Letters*, 88(10):108701, 2002.

[23] X. Cheng, H. Wang, and Q. Ouyang. Scale-free network model of node and connection diversity. *Physical Review E*, 65:066115, 2002.

[24] A. Wagner. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, 17(12):1183–1197, 2001.

[25] A. Bhan, D. J. Galas, and T. G. Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.

[26] A. V. Lukashin, M. E. Lukashev, and R. Fuchs. Topology of gene expression networks as revealed by data mining and modeling. *Bioinformatics*, 19(15):1909–1916, 2003.

[27] M. J. Gagen and J. S. Mattick. Failed "nonaccelerating" models of prokaryote gene regulatory networks. 2003. arXive:q-bio.MN/0312022. See http://arxiv.org/abs/q-bio.MN/0312022.

[28] I. Cases, V. de Lorenzo, and C. A. Ouzounis. Transcription regulation and environmental adaptation in bacteria. *Trends in Microbiology*, 11(6):248–253, 2003.

[29] C. K. Stover, et al. Complete genome sequence of *Pseu-*

domonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 406(6799):959–964, 2000.

[30] S. D. Bentley, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, 417(6885):141–147, 2002.

[31] J. L. Cherry. Genome size and operon content. *Journal of Theoretical Biology*, 221:401–410, 2003.

[32] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31:64–68, 2002.

[33] D. Thieffry. From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in *Escherichia coli*. *BioEssays*, 20(5):433–440, 1998.

[34] M. Madan Babu and S. A. Teichmann. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Research*, 31(4):1234–1244, 2003.

[35] I. Yanai, C. J. Camacho, and C. DeLisi. Predictions of gene family distributions in microbial genomes: Evolution by gene duplication and modification. *Physical Review Letters*, 85(12):2641–2644, January 2000.

[36] V. Kunin and C. A. Ouzounis. The balance of driving forces during genome evolution in Prokaryotes. *Genome Research*, 13:1589–1594, 2003.

[37] A. L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272(1-2):173–187, 1999.

[38] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[39] S. N. Dorogovtsev and J. F. F. Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, 63:056125, 2001.

[40] N. Rosenfeld, M. B. Elowitz, and U. Alon. Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology*, 323:785–793, 2002.

[41] P. B. Warren and P. R. ten Wolde. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. q-bio.MN/0310029 (http://arxiv.org/abs/q-bio.MN/0310029), 2003.

[42] P. Erdös and A. Renyi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, 5:17–61, 1960.

[43] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402 (Supp):C47–C52, 1999.

[44] J. Vogel, V. Bartels, T. H. Tang, G. Churakov, J. G. Slagter-Jäger, A. Hüttenhofer, and E. G. H. Wagner. RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Research*, 31(22):6435–6443, 2003.

[45] J. S. Mattick and M. J. Gagen. The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Molecular Biology and Evolution*, 18(9):1611–1630, 2001.

[46] J. S. Mattick. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Reports*, 2(11):986–991, 2001.

[47] J. S. Mattick. Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays*, 25:930–939, 2003.